The AANVIS project: towards the automatic multilingual analysis of non-verbal information in speech

Daniel J. Hirst, Saandia Ali, Amina Chentir, Hyongsil Cho, Irina Nesterenko, Stéphane Rauzy

ICPhS 2007 Satellite Meeting
Workshop on Intonational Phonology: Understudied or Fieldwork Languages
August 5, 2007. Saarbrücken, Germany

Presentation of the Aanvis project: towards the automatic analysis of non-verbal information in speech.

Daniel Hirst, Saandia Ali, Amina Chentir, Hyongsil Cho, Irina Nesterenko, Stéphane Rauzy.

The aim of this project is to develop an objective and empirical methodology for the analysis of non-verbal information obtained from speech corpora of several languages representative of different prosodic types. This will lead to obtaining fundamental knowledge about the prosodic systems of these languages that will be directly applicable to the automatic multilingual processing and interpretation of the oral language.

With the rapid growth of international trade and travel, people are increasingly finding themselves in contact with people who do not speak a common language. Although speech-to-speech translation has become a viable technology as a language education tool or means of interactive communication, the technology needs to be further refined as it currently is unable to take into account non-verbal information such as the differences in the tone of voice and speech style which can carry information crucial to a proper understanding of speech.

The natural barrier between speech communities is undoubtedly one of the major problems with which modern multicultural societies are confronted. Speech technology continues to contribute its share, but the analysis of non-verbal information has made comparatively little progress, leaving almost unknown the multicultural aspects of various speech styles.

Speech technology today is far behind the level of sophistication of the automatic processing of written language. One must note that the majority of applications of these technologies do not make use of the potential specificities of spontaneous speech, in particular of speech prosody, which plays a central role in oral communication. The majority of speech synthesis systems are based on rudimentary prosody, which is one main reason why these systems are not perceived as acceptable replacements for human operators. In the same way, contrary to human beings, current automatic speech recognition systems take little or no account of most prosodic information in the interpretation of the utterances.

We propose in this project to combine the expertise from a number of different fields. Partners will include:

- specialists in speech science and technology
- specialists on the prosody of specific languages
- specialists on the modelling of speech prosody on a cross-linguistic basis
- specialists on natural language processing.

We will develop a multilingual framework for the data processing of the non-verbal information of speech that will allow an interpretation of speakers' intentions as well as the linguistic contents of the utterances, for a better comprehension of not only "what was said" but also "how it was meant to be interpreted". This would considerably reduce potential ambiguity in speech translation and provide useful technologies for dictation applications and speech translation.

One of the basic assumptions behind this project is that an efficient way to model the relationship between linguistic meanings and linguistic forms is by establishing a clear distinction between functional representations (eoncoding meanings) and formal representations (encoding sounds).

The ultimate aim is to predict functions from forms (= interpretation) but the methodology proposed in this project will consist in first predicting forms from a limited set of well-established functions, then in classifying the set of variants forms observed with each function and finally giving functional labels to the observed variant forms.

We thus intend to proceed by an incremental system of analysis by synthesis, gradually enriching the labels of the functional representation until we are able to predict a satisfactory formal representation of the corpus.

For each language we will obtain or constitute a version of the continuous passages of the Eurom1 corpus developed during the European Esprit project SAM (Speech Assessment and Methodology). This corpus, which already exists for a large number of languages, will be recorded by a representative number of speakers (eg at least 5 male and 5 female). The corpus will be analysed using tools for the automatic analysis of prosody developed in Aix en Provence, some of which have already been used in the European Multext project. The results of the analysis will make it possible to define a preliminary set of prosodic patterns for each language. A symbolic coding of the prosodic forms will then be put into relationship with a symbolic representation of prosodic functions by an incremental system of analysis by synthesis (Hirst 2005, Hirst & Auran 2005) using state of the art techniques of natural language processing.

The representations obtained from the first corpus will then be applied to larger corpora of more authentic speech, compatible with the various different applications envisaged. It is anticipated that by using the same base corpus for each language, the initial range of prosodic patterns will be constrained to a comparable subset for the different languages which will make it possible to establish direct comparisons of the relations between prosodic forms and prosodic functions across languages. The same methodology is being applied to a number of different languages with very different typological characteristics. Currently, work has begun or is about to begin on English, French, Italian, German, Brazilian Portuguese, Russian, Arabic, Finnish, Chinese and Korean.

In this presentation we will present preliminary results for the methodology as applied to French, English, Russian, Arabic and Korean.

References

[1] Hirst, D.J. 2005. Form and function in the representation of speech prosody. in K.Hirose, D.J.Hirst & Y.Sagisaka (eds) *Quantitative prosody modeling for natural speech description and generation* (=*Speech Communication* 46 (3-4)), 334-347

[2] Hirst, D.J. & Auran, C. 2005. Analysis by synthesis of speech prosody. The ProZed environment. in *Proceedings of Eurospeech/Interspeech* Lisbonne octobre 2005.